# The Comparative Method Rules! Codon Volatility Cannot Detect Positive Darwinian Selection Using a Single Genome Sequence

*Tal Dagan* and Dan Graur†*

*Department of Zoology, George S. Wise Faculty of Life Science, Tel Aviv University, Ramat Aviv, Israel; and †Department of Biology and Biochemistry, University of Houston, Houston, Texas

All established methods for detecting positive selection at the molecular level rely on comparisons between nucleotide sequences. An exceptional method that purports to detect selection on the basis of a single genomic sequence has recently been proposed. This method uses a measure called "codon volatility," defined for each codon as the ratio between the number of nonsynonymous codons that differ from the codon under study at a single nucleotide position and the number of sense codons that differ from the codon under study at a single nucleotide position. Here, we examine various properties of codon volatility and its derivatives and use simulation of evolutionary processes to determine whether they can be used to detect selective pressures. Codons for only four amino acids (glycine, leucine, arginine, and serine) show any variation in codon volatility. Thus, codon volatility is mainly a proxy for amino acid usage, rather than for codon usage, with 65% of all synonymous changes and 27% of all nonsynonymous changes being undetectable by this measure. Genes identified by the volatility method as being subject to positive selection tend to have idiosyncratic amino acid compositions (e.g., they are glycine rich or arginine poor). An additional property of codon volatility is the near zero variance of its mean expectation, which translates into overestimated statistical significance estimates, especially in the absence of corrections for multiple comparisons. A comparison with measures of selection inferred through comparative methodology reveals no relationship between the results of the two methods. Finally, we show that codon volatility can increase in the absence of positive Darwinian selection; that is, increased codon volatility is not indicative of positive selection.

## Introduction

All established methods for detecting positive selection at the molecular level (as well as all other molecular evolutionary methods) rely on comparisons between nucleotide sequences. A method that purports to detect positive (as well as negative) selection on the basis of a single genomic sequence has recently been proposed (Plotkin, Dushoff, and Fraser 2004). This method, which we shall henceforth refer to as PDF2004, is revolutionary not only because it purports to require "far fewer data than comparative sequence analysis" but also because it challenges the essentiality of the comparative method in at least one area of evolutionary research.

PDF2004 is based on a measure called "codon volatility." Several variants of codon volatility have been proposed by Plotkin and Dushoff (2003), but PDF2004 uses the simplest variant, in which the volatility of a codon, $v(C)$, is defined as

$$v(C) = \frac{N}{T} \quad (1)$$

where $N$ is the number of nonsynonymous codons that differ from codon $C$ at a single nucleotide position, and $T$ is the number of sense (nontermination) codons that differ from codon $C$ at a single nucleotide position. As an illustration, let us consider the nine neighbor codons of AGA; that is, codons that differ from AGA (Arg) at one nucleotide position. Two neighbor codons are synonymous (CGA and AGG), six are nonsynonymous (GGA, AGT, AGC, AAA, ACA, and ATA), and one is a stop codon (TGA). The volatility of codon AGA is, therefore, $6/8 = 0.75$. The volatility of a gene, $v(G)$, was defined by Plotkin, Dushoff, and Fraser (2004) as the sum of the volatilities of its constituent codons.

The PDF2004 method for detecting positive selection relies on the premise that high gene volatility is indicative of an excess of amino acid replacements and, hence, of positive selection (at least in the recent evolutionary past). Under this paradigm, a genome can be scanned, and genes with observed volatility values that are significantly higher than their expected volatility are deemed to have been subject to positive Darwinian selection. Conversely, genes exhibiting exceptionally low volatilities are deemed to have been subjected to strong purifying selection.

The PDF2004 method was applied to two completely sequenced genomes: a bacterium (*Mycobacterium tuberculosis*) and a unicellular eukaryote (*Plasmodium falciparum*). Based on these analyses, 9% to 10% of all protein-coding genes were found to be significantly more volatile than the genome as a whole ($P < 0.05$); that is, they may be said to have evolved under positive selection. These values are much larger than estimates obtained through traditional comparative analyses (e.g., Endo, Ikeo, and Gojobori 1996). The volatility results indicative of purifying selection are even more peculiar. Only 7.5% of the protein-coding genes in *M. tuberculosis* and 13.9% of the protein-coding genes in *P. falciparum* exhibit any telltale signs of purifying selection, whereas the current knowledge in evolutionary biology is that virtually all protein-coding genes are subject to purifying selection.

In this note, we examine the PDF2004 method in some detail to determine whether it can be used to detect positive Darwinian selection.

## Materials and Methods

Our study was divided into two parts. In the first part, we examined various properties of codon volatility, which

Key words: Codon volatility, comparative method, positive selection, synonymous and nonsynonymous substitutions, Darwinian purifying selection.

E-mail: dgraur@uh.edu.

*Mol. Biol. Evol.* 22(3):496–500. 2004
doi:10.1093/molbev/msi033
Advance Access publication November 3, 2004

are conspicuously absent from the original descriptions (Plotkin and Dushoff 2003; Plotkin, Dushoff, and Fraser 2004). In the second part, we studied four evolutionary scenarios by using whole genomic-coding sequences as start points in computer simulations and asked whether increases in codon-volatility values are exclusively indicative of positive selection.

### Codon Usage and Genome Volatility

The frequency of each codon (excluding stop codons) was counted for all the protein-coding genes in 171 completely sequenced genomes of bacteria from the July 2004 version of the NCBI database (ftp://ftp.ncbi.nlm.nih.gov/). For each genome, preferred codon usage for each amino acid was defined as the codon with the highest frequency from among the different codons for the same amino acid. The universal genetic code was used to calculate codon volatility. For each genome, mean genome volatility was calculated as the mean volatility of all codons weighted by their frequency within the genome. Mean genome volatility of preferred codons was calculated as the mean volatility of the preferred codons weighted by the frequency of their respective amino acids within the proteome.

### Simulated Evolution

To test the effect of different selection regimes on the predictions made by the volatility method, we simulated evolutionary processes for all the genes within three genomes that were chosen because they exhibit low, intermediate, and high mean volatilities. The mean volatilities of *Mycobacterium tuberculosis*, *Geobacter sulfurreducens*, and *Clostridium acetobutylicum* are 0.733, 0.746, 0.783, respectively.

The parameters for the simulation included (1) numbers of point mutations, (2) transition/transversion ratios, and (3) fixation probabilities for two types of nonsynonymous mutation and three types of synonymous mutation (see below). In each step of simulation, one of the nucleotides within a gene was mutated. This mutation was then eliminated or fixed according to the specified selectional regime. That is, for each gene, a site was chosen randomly from a uniform distribution ranging between 1 and the length of the gene. The nucleotide at the site was then randomly mutated into another according to the predetermined transition/transversion ratio. The mutation was then classified into synonymous, nonsynonymous, or nonsense. Synonymous mutations were further classified into (1) preferred codon to nonpreferred codon, (2) nonpreferred codon to preferred codon, or (3) nonpreferred codon to another nonpreferred codon. Amino acid replacements resulting from nonsynonymous mutations were further classified into radical or conservative replacements, depending on whether the Grantham (1974) physicochemical distance between the exchanged amino acids was larger or smaller than 100, respectively.

Four selective regimes were simulated (table 1). In the (completely) neutral regime, all sense and missense mutations were fixed. In the purifying selection regime, nonsynonymous mutations were selected against. The probability of fixation for a radical amino acid replacement was smaller than that for a conservative replacement. Synonymous mutations to nonpreferred codons were also selected against, but to a lesser extent than nonsynonymous mutations. In the so-called synonymous regime, nonsynonymous mutations were invariably eliminated, and only synonymous mutations were allowed to fix. Synonymous mutations to preferred codons were assigned fixation probabilities of 1. Synonymous mutations to nonpreferred codons were assigned fixation probabilities of 0.5. Under the so-called nonsynonymous regime, all of the nonsynonymous mutations and about 50% of the synonymous mutations were fixed. This regime is supposed to simulate an excess of nonsynonymous over synonymous substitutions as expected under positive selection. A mutation was fixed if the value drawn from a uniform distribution between 0 and 1 was smaller than the fixation parameter for the type of mutation in the selectional regime in question.

For each evolutionary regime, we simulated 50, 100, 150, 200, 500, 1,000, 1,500, 2,000, and 10,000 generations. We simulated two transition/transversion ratios: 1:1 and 5:1. (The transition/transversion ratio turned out to have very little influence on the results and, hence, we only present results derived from the 1:1 transition/transversion simulations.)

Gene volatilities and corresponding $P$ values were calculated using the PDF2004 software at http://www.cgr.harvard.edu/volatility.

### Comparison of the PDF2004 Method with the Comparative Method

The results of the PDF2004 method were compared with those obtained from a comparison between the genomes of *M. tuberculosis* (strain CDC1551) and *M. bovis*. We identified putative orthologs by the method of reciprocal best Blast hits with a cutoff of $e < 0.001$. Of the 4,099 *M. tuberculosis* genes, 3,723 orthologs were identified in the genome of *M. bovis*. The sequences of the orthologous genes were aligned to the corresponding protein sequence of *M. tuberculosis* using Wise2 (Birney, Clamp, and Durbin 2004). This alignment respects codon positions as needed for the calculation of synonymous and nonsynonymous rates of substitution. Rates of synonymous (dS) and nonsynonymous (dN) substitutions were estimated using PAML (Yang 1997). We tested for correlation between the results of PDF2004 and the comparative method (dN/dS) by using Spearman rank nonparametric correlation (Zar 1999).

## Results and Discussion

As seen in table 2, codon volatility in the universal genetic code ranges between 0.5 in CGA (Arg) and 1 in ATG (Met) and TGG (Trp). In fact, codon volatility may only assume 12 possible values: 0.5, 0.56, 0.57, 0.63, 0.67, 0.71, 0.75, 0.78, 0.86, 0.88, 0.89, and 1. The most common codon volatility value is 0.67 for 25 codons specifying Ala, Gly, Leu, Pro, Arg, Ser, Thr, and Val. The

**Table 1**
**Fixation Probabilities in Four Simulated Selection Regimes**

| | Nonsynonymous Substitution | | Synonymous Substitutions | | | |
|---|---|---|---|---|---|---|
| Regime | Radical | Conservative | Preferred Codon to Nonpreferred Codon | Nonpreferred Codon to Preferred Codon | Nonpreferred Codon to Nonpreferred Codon | Nonsense |
| Neutral | 1 | 1 | 1 | 1 | 1 | 0 |
| Purifying selection | 0.2 | 0.5 | 0.8 | 1 | 0.8 | 0 |
| Synonymous | 0 | 0 | 0.5 | 1 | 0.5 | 0 |
| Nonsynonymous | 1 | 1 | 0.5 | 0.5 | 0.5 | 0 |

second and third most common codon volatility values are 0.89 for 10 codons encoding Asp, Phe, His, Asn, and Ser, and 0.88 for eight codons specifying Cys, Glu, Lys, and Gln. Sixteen out of the 20 standard amino acids are each specified by codons with identical volatilities. For instance, all four codons for Pro have the same volatility. Codons for only four amino acids (Gly, Leu, Arg, and Ser) show any variation in codon volatility. Thus, codon volatility turns out to be mainly a proxy for amino acid usage, rather than for codon usage. Given that different genes have different amino acid requirements, and given that gene volatility measures mostly amino acid composition, it would be difficult on a priori grounds to see how an exceptional amino acid composition of a certain gene may be used as a measure of selective pressure. In fact, the PE and PPE families of genes in *M. tuberculosis*, which were singled out in PDF2004 as extreme instances of evolution under positive selection, are only exceptional in their amino acid composition in comparison to the rest of the *M. tuberculosis* genes. These genes are extremely rich or extremely poor in amino acids that contribute to the variation in volatility values. The PE and PPE genes are very rich in glycine (Tekaia et al. 1999; Kinsella et al. 2003). In *P. falciparum*, the genes identified as showing "the strongest signs of positive selection" are only exceptional in their amino acid composition by having unusually low levels of arginine.

In PDF2004, the statistical significance of an observed gene-volatility value was determined by comparison to the "expected volatility of the gene." In theory, the expected volatility was supposed to be calculated by a bootstrap distribution of $10^6$ synonymous versions of the gene, in which each version is a new sequence that is created with the same amino acid usage as the gene, and a codon usage that is drawn randomly according to the codon usage of the genome. At this high number of repeats, the expected volatility turned out to be calculable directly from the codon usage of the genome. Plotkin et al. (2004) state: "We calculate the volatility $P$ value for G by comparing the gene's observed volatility to its expected volatility, given the amino acid content of the gene and the codon usage in the entire genome."

The expected volatility and variance is computed for each amino acid as

$$E[v(a)] = \sum_{i \in codons} v(i) \frac{N_i}{M_a} \quad (2)$$

$$V[v(a)] = \sum_{i \in codons} v(a)^2 \frac{N_i}{M_a} - E[v(a)]^2 \quad (3)$$

where $v(i)$ is the codon volatility of codon $i$, $N_i$ is the number of occurrences of codon $i$ in the genome, and $M_a$ is the number of occurrences of amino acid $a$ in the proteome.

We note that for the 16 amino acids with invariant codon volatilities (i.e., any amino acid that is not Gly, Leu, Arg, or Ser) the expected amino acid volatility is independent of either codon or amino acid frequencies:

**Table 2**
**Volatilities of Codons in the Universal Genetic Code**

| First Position (5′ end) | Second Position | | | | Third Position (3′ end) |
|---|---|---|---|---|---|
| T | T | C | A | G | |
| | Phe (0.89) | Ser (0.67) | Tyr (0.86) | Cys (0.88) | T |
| | Phe (0.89) | Ser (0.67) | Tyr (0.86) | Cys (0.88) | C |
| | Leu (0.71) | Ser (0.57) | STOP | STOP | A |
| | Leu (0.75) | Ser (0.63) | STOP | Trp (1.00) | G |
| C | Leu (0.67) | Pro (0.67) | His (0.89) | Arg (0.67) | T |
| | Leu (0.67) | Pro (0.67) | His (0.89) | Arg (0.67) | C |
| | Leu (0.56) | Pro (0.67) | Gln (0.88) | Arg (0.50) | A |
| | Leu (0.56) | Pro (0.67) | Gln (0.88) | Arg (0.56) | G |
| A | Ile (0.78) | Thr (0.67) | Asn (0.89) | Ser (0.89) | T |
| | Ile (0.78) | Thr (0.67) | Asn (0.89) | Ser (0.89) | C |
| | Ile (0.78) | Thr (0.67) | Lys (0.88) | Arg (0.75) | A |
| | Met (1.00) | Thr (0.67) | Lys (0.88) | Arg (0.78) | G |
| G | Val (0.67) | Ala (0.67) | Asp (0.89) | Gly (0.67) | T |
| | Val (0.67) | Ala (0.67) | Asp (0.89) | Gly (0.67) | C |
| | Val (0.67) | Ala (0.67) | Glu (0.88) | Gly (0.63) | A |
| | Val (0.67) | Ala (0.67) | Glu (0.88) | Gly (0.67) | G |

NOTE.—Amino acids whose codons exhibit variation in volatility are shaded.

$$E[v(a)] = v(a) \quad (4)$$

where $v(a)$ is the volatility of any codon encoding the amino acid.

What is, however, more important is that for these 16 amino acids, the variance of the expected volatility is zero. The variances for the remaining amino acids are close to zero: 0.0061, 0.00015, 0.0053, and 0.014 for the Gly, Leu, Arg, and Ser, respectively.

The expected volatility of a gene (G) and its variance are calculated as

$$E[v(G)] = \sum_{a \in a.a.} m_a E[v(a)] \quad (5)$$

$$V[v(G)] = \sum_{a \in a.a.} m_a V[v(a)] \quad (6)$$

where $m_a$ is the amino acid usage for amino acid $a$ in gene $G$. Because the variances of most amino acids are zero, and those of Gly, Leu, Arg, and Ser are close to zero, the expected volatility variance of a gene should be very small. Indeed, the coefficients of variance of the expected volatilities of *M. tubeculosis* genes range between 0.03% and 0.7%.

As stated previously, the volatility $P$ value of a gene is calculated by comparing its observed volatility to the expected volatility. The expected volatility typically has a tiny variance, hence, it is easy to overestimate the statistical significance of the observed values. Moreover, the calculation of $P$ values in the PDF2004 method involves thousands of dependent comparisons, and, therefore, a Bonferroni correction for multiple comparisons should be used in assessing statistical significance. Thus, in the case of the 4,099 genes from *M. tuberculosis*, $P$ values should be smaller than $0.05/4{,}099 = 1.21 \times 10^{-5}$to be significant. Without Bonferroni correction, 376 *M. tuberculosis* genes have $P$ values smaller than 0.05. With the Bonferroni correction, the number of genes associated with statistically significant $P$ values is reduced to 14. The corresponding numbers in *P. falciparum* are 534 and 52, respectively.

As we have shown previously, gene volatility is determined solely by 22 codons encoding four amino acids. Given the very low variation in codon volatility, it is reasonable to assume that even a small excess of high-volatility codons in comparison with their frequency in the genome may result in statistically significant volatility $P$ values. To address this issue, we tested the correlation between the frequency of the 22 codons specifying Gly, Leu, Arg, and Ser in *M. tuberculosis* and the volatility $P$ values. We also used multiple correlation to assess how much of the variation in $P$ values can be explained by combinations of codons (table 3). Surprisingly, we found that with only two codons (CTG and AGC), we can explain about 42% of the variation. Using all 22 codons for Gly, Leu, Arg, and Se, we can explain up to 70% of the variation in volatility $P$ value.

We calculated mean volatility of preferred codon usage and all codons in 171 bacterial genomes. Genome volatility of preferred codons ranges between 0.724 in *Streptomyces coelicolor* and 0.800 in *Mycoplasma genitalium*, with a mean of 0.765 ± 0.019. Whole-genome volatility ranges between 0.728 in *Streptomyces coelicolor* and 0.79 in *Wigglesworthia glossinidia*, with mean of 0.765 ± 0.015 (table S1 in Supplemental Material online). The difference between the mean volatility of preferred codons and all codons is −0.0008 ± 0.0077. There are 79 bacteria in which whole-genome volatility was larger than the volatility for preferred codons and 92 cases in which the opposite was true. The genome of *Mycobacterium tuberculosis* that was used as a case study in Plotkin, Dushoff, and Fraser (2004) exhibits a low volatility level with a mean volatility of preferred codons of 0.726 and mean volatility of the whole genome of 0.733.

**Table 3**
**Correlation Coefficient (r) of Volatility *P* Values and Codon Usages for Codons of Amino Acids with Heterogeneous Codon Volatilities**

| Codon | Amino Acid | r | Cumulative $R^2$ |
|---|---|---|---|
| CTG | Leu | 0.52 | 0.228 |
| AGC | Ser | −0.46 | 0.418 |
| TTG | Leu | −0.38 | 0.455 |
| TCG | Ser | 0.37 | 0.474 |
| CGG | Arg | 0.3 | 0.523 |
| AGT | Ser | −0.23 | 0.581 |
| AGG | Ser | −0.2 | 0.601 |
| TCC | Ser | 0.16 | 0.606 |
| CTC | Leu | −0.16 | 0.608 |
| TCA | Ser | 0.14 | 0.615 |
| AGA | Arg | −0.14 | 0.621 |
| CTT | Leu | −0.14 | 0.628 |
| CGA | Arg | −0.14 | 0.686 |
| CGC | Arg | −0.13 | 0.687 |
| TTA | Leu | −0.09 | 0.698 |
| CTA | Leu | −0.07 | 0.699 |
| CGC | Arg | −0.06 | 0.7 |
| GGT | Gly | −0.06 | 0.701 |
| GGA | Gly | 0.06 | 0.703 |
| GGC | Gly | 0.04 | 0.704 |
| TCT | Ser | 0.03 | 0.704 |
| GGG | Gly | 0.005 | 0.704 |

NOTE.—The cumulative explained variability ($R^2$) was calculated in a stepwise manner, in which the variables (codons) are added according to their correlation coefficient in descending order (Zar 1999).

The results of the simulations (table 4, and see figures F1–F3 in Supplementary Material online) indicate that regardless of initial codon composition and selection regime, mean gene volatility tends to approach asymptotically an equilibrium value. If the initial volatility is above the equilibrium value, it will decrease during evolution; if it is below the equilibrium value, it will increase. The exact value at equilibrium will obviously be determined by the amino acid frequencies and codon usage. In particular, gene volatility may be affected by whether the codon usage is biased towards low-volatility or high-volatility codons. In the simplest case, in which all 61 codons are equally frequent, the mean volatility will be approximately 0.747.

What is, however, the most important thing to notice from our simulations, is that volatility may increase in the absence of positive selection. Conversely, it may decrease in the absence of purifying selection. For example, in our simulations with the *M. tuberculosis* coding sequences as starting points, volatility went up from 0.735 to 0.749

**Table 4**
**Mean Volatility in the Simulated Generations of the Three Bacterial Genomes**

| | *Mycobacterium tuberculosis* | | | | *Geobacter sulfurreducens* | | | | *Clostridium acetobutylicum* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Number of Mutations | Purifying Selection | Neutral | Synonymous | Nonsynonymous | Purifying Selection | Neutral | Synonymous | Nonsynonymous | Purifying Selection | Neutral | Synonymous | Nonsynonymous |
| 0 | 0.735 | 0.735 | 0.735 | 0.735 | 0.750 | 0.750 | 0.750 | 0.750 | 0.787 | 0.787 | 0.787 | 0.787 |
| 50 | 0.736 | 0.737 | 0.736 | 0.737 | 0.749 | 0.750 | 0.749 | 0.750 | 0.785 | 0.782 | 0.787 | 0.782 |
| 100 | 0.737 | 0.739 | 0.736 | 0.739 | 0.749 | 0.750 | 0.749 | 0.750 | 0.783 | 0.778 | 0.787 | 0.778 |
| 150 | 0.738 | 0.740 | 0.736 | 0.740 | 0.749 | 0.750 | 0.749 | 0.750 | 0.781 | 0.775 | 0.786 | 0.775 |
| 200 | 0.738 | 0.741 | 0.736 | 0.741 | 0.749 | 0.750 | 0.749 | 0.750 | 0.779 | 0.772 | 0.786 | 0.772 |
| 500 | 0.741 | 0.744 | 0.736 | 0.744 | 0.749 | 0.750 | 0.749 | 0.751 | 0.772 | 0.762 | 0.786 | 0.762 |
| 1000 | 0.743 | 0.747 | 0.737 | 0.747 | 0.749 | 0.750 | 0.749 | 0.751 | 0.765 | 0.756 | 0.785 | 0.755 |
| 1500 | 0.744 | 0.748 | 0.737 | 0.748 | 0.749 | 0.750 | 0.749 | 0.750 | 0.761 | 0.753 | 0.785 | 0.752 |
| 2000 | 0.745 | 0.748 | 0.738 | 0.749 | 0.749 | 0.750 | 0.750 | 0.750 | 0.758 | 0.751 | 0.785 | 0.751 |
| 10000 | 0.748 | 0.749 | 0.739 | 0.749 | 0.749 | 0.749 | 0.751 | 0.749 | 0.749 | 0.749 | 0.783 | 0.749 |

under the completely neutral evolution regime. Moreover, under the nonsynonymous regime, which was intended to emulate positive selection, volatility went down from 0.787 to 0.749 in the simulations with *C. acetobutylicum* coding sequences as starting points.

We compared the results of the PDF2004 method with the dN/dS ratios obtained by comparing orthologous genes from *M. tuberculosis* and *M. bovis*. There were no significant correlation coefficients between either expected or observed volatility in *M. tuberculosis* on the one hand and dN, dS, and dN/dS on the other. As far as the volatility *P* values are concerned, we obtained no significant correlations with either dN or dN/dS. A tiny negative correlation was seen between volatility *P* values and dS ($r^2$=0.0013). To illustrate this point, we note that in several cases, orthologous genes may be identical in sequence between *M. tuberculosis* and *M. bovis*, and yet they may show volatility *P* values indicative of positive selection. One such case is gene MT0441 (a hypothetical membrane protein in *M. tuberculosis*), which is identical in sequence to its ortholog in *M. bovis* but has volatility *P* value of $8.97 \times 10^{-6}$. Thus, the comparative method indicates very strong purifying selection, whereas the PDF2004 method "reveals" an instance of strong positive Darwinian selection that paradoxically results in no evolutionary changes in either gene. We must, therefore, conclude that the PDF2004 method cannot detect selective pressures.

## Acknowledgments

We thank J. B. Plotkin for helpful feedback. T.D. was supported in part by a scholarship in Complexity Science from the Yeshaia Horvitz Association.

## Literature Cited

Birney E., M. Clamp, and R. Durbin. 2004. Genewise and genomewise. Genome Res. **14**:988–995.

Endo, T., K. Ikeo, and T. Gojobori. 1996. Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol. **13**:685–690.

Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. Science **85**:862–864.

Kinsella R. J., D. A. Fitzpatrick, C. J. Creevey, and J. O. Mclnerney. 2003. Fatty acid biosythesis in *Mycobacterium tubeculosis*: lateral gene transfer, adaptive evolution, and gene duplication. Proc. Natl. Acad. Sci. USA **100**:10320–10325.

Plotkin, J. B., and J. Dushoff. 2003. Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. Proc. Natl. Acad. Sci. USA **100**:7152–7157.

Plotkin, J. B., J. Dushoff, and H. B. Fraser. 2004. Detecting selection using a *M. tuberculosis* and *P. falciparum*. Nature **428**:942–945.

Tekaia F., S. V. Gordon, T. Garnier, R. Brosch, B. G. Barrell, and S. T. Cole 1999. Analysis of the proteome of *Mycobacterium tuberculosis* in silico. Tuber. Lung Dis. **79**:329–342.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS **13**:555–556.

Zar, J. H. 1999. Biostatistical analysis. Prentice Hall, Upper Saddle River, NJ.

William Martin, Associate Editor

Accepted October 26, 2004